International Journal of Engineering & Technology, 7 (1.3) (2018) 66-68

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Potential item set mining by using utility pattern growth model in big data

S. Angel Latha Mary *, R. Divya, K. Uma Maheswari

Karpagam College of Engineering
** Corresponding author E-mail: xavierangellatha@gmail.com*

Abstract

Information extracted by using data mining in earlier days. Now a day's, the most talked about technology is Big Data. Utility Mining is the most crucial task in the real time application where the customers prefer to choose the item set which can yield more profit. Handling of large volume of transactional patterns becomes the complex issue in every application which is resolved in the existing work introducing the parallel utility mining process which will process the candidate item sets in the paralyzed manner by dividing the entire tasks into sub partition. Each sub partition would be processed in individual mapper and then be resulted with the final output value. The time complexity would be more when processing an unnecessary candidate item sets. This problem is resolved in the proposed methodology by introducing the novel approach called UP-Growth and UP-Growth+ which will prune the candidate item sets to reduce the dimension of the candidate item sets. The time complexity is further reduced by representing the candidate item sets in the tree layout. The test results prove that the proposed new approach provides better result than the existing work in terms of accuracy.

Keywords: *High Utility Itemsets, UP-Tree, Utility Mining.*

1. Introduction

The purpose of utility mining is to require out valued and supportive data from knowledge by taking into account of response, capability, price or different customer preferences. High utility itemset (HUI) mining is one of the significant task in utility mining, which can be used to determine sets of items bring high utilities (e.g., high profits). This data has been applied to several applications such as open market place analysis, transportable computing and flat bioinformatics. Because of its spacious range of applications, many studies proposed for mining HUIs in databases. The majority of them deduce that data are stored in federal database with a separate machine performing the mining responsibilities. However, in large data environments, data may be originated from ambiguous sources and extremely circulated. In this system, this paper considers the above challenges, and proposes a replacement structure for mining high helpfulness itemsets in big data. This paper first recommend work of narrative algorithm named Parallel mining High Utility Itemsets by pattern-Growth (PHUI-Growth) implemented on a Hadoop platform for instance simple operation in high level language, fault tolerance, low announcement overheads and high scalability on product hardware. Second PHUI-Growth adopts the MapReduce structural design to divide the entire mining task into minor modules and uses Hadoop Distributed File System (HDFS) to process circulated data. Thus, it can equal mining of HUIs from distributed databases and crossways multiple service computers in a consistent method. Third, PHUI-Growth adopts Discarding local unpromising items in MapReduce framework (DLU-MR) to successfully prune the search space and unwanted middle itemsets formed during the mining method, which further increases the performance of PHUI-Growth. The main problem occurs in the utility set mining is the time complication and memory storage problem where the large volume of transactional patterns are present. The larger number of transaction patterns might leads to the performance degradation while extracting the high utility item sets.To improve the performance this paper uses compact tree arrangement, named UP-Tree, to preserve the information of transactions and avoid scanning database more than one time. Ying Chun Lin et al.[1], are proposed two strategies to lessen the overestimated utilities stored in the nodes of global UP-Tree which receives promising items and discouraging items. After receiving all promising items, DGU is applied and all transactions are updated Any ordering can be used such as the lexicographical, support, or TWU order used by deleting the unnecessary items and sorting the outstanding promising items in a fixed order.

2. Material and methods

A. Haui mining

Pattern mining plays a significant role in data mining. Frequent pattern mining searches the associations and mutual relationship between items in transactional or relational datasets. Chan suggested utility mining to solve by deleting the unnecessary items and sorting the outstanding promising items in a fixed order used a tree structure to extract high utility itemsets (HUIs) of which utility standards are larger than or equal a predefined threshold. Utility mining is based on the value of itemset. This average utility measure was recommended by Hong [6] where it considered the length of itemset. If an itemset which is not a HAUI combined with other items and to be converted into HAUI. This novel method generat-

ed candidates more rapidly and compared with threshold to obtain HAUIs with their average utility standards. This approach is similar to WIT-tree[7] in which is used for mining frequent itemsets. Some other methods have been anticipated to prune candidates and accumulate time. This technique is to condense candidates proficiently by means of HAUI-Tree and to mine HAUI from transaction databases.

B. Two-phase algorithm

It professionally trim down the number of candidates and be capable of accurately acquire the high utility itemsets. In the primary stage, a representation that applies the "transaction-weighted downward closure proprietary" which can be used to enhance the performance by reduce the number of candidate of items. In the subsequent stage, one more database scan is performed to identify or classify the high utility item sets. Also parallelize the algorithm was suggested by Liu .Y et al. [4], proposed on shared memory multi-process structural design using regular Count Partitioned Database (CCPD) approach by making an allowance for the different values of individual things as utilities utility withdrawal focuses on identifying the itemsets with high profit. Because "downward closure property" doesn't concern to utility mining, the FORMATION of candidate itemsets is mainly valuable in terms of time and storage space.

C. Thui-tree algorithm

This algorithm requires more than one database scan and doesn't meet requirements of data stream. It cannot be applied for landmark and time fading window models. To overcome this drawback another two algorithms MHUI-BIT and MHUI-TID are proposed by representing data in the form of Bitvecot(BIT) and Tidlist(TID). However, MHUI-TID and MHUI-BIT achieved efficiency in terms of memory consumption and execution time over THUI-Mine. But these two algorithms required more than one database scan due to level-wise approach[2].

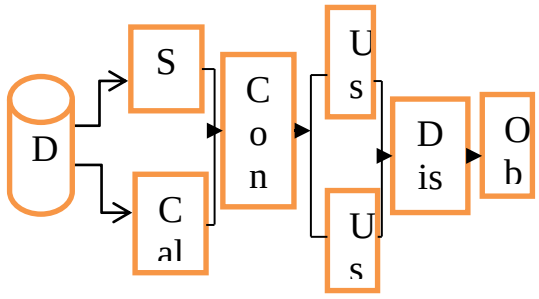

Fig. 1: Implementation of UP-Growth HighUtility Itemsets Algorithm

The figure 1 describe the structural design of the future system. The input data set is divided into required number of partitions. The partitions are send to the mapper phase which is used to compute k clusters for each partition of the dataset and the local clusters are obtained. Next in the reducer phase the clusters are merged to obtain the final resultant clusters.

3. Results and discussion

In the proposed system the process involves three phases. The first stage involves counting the input dataset and partitioning. The second stage involves implementing map function and then implementing the reduce function.

A. Counting phase

The input database DD can be viewed as a set of transactions that are stored in several computers. In counting phase, the algorithm takes one MapReduce pass to parallel counts TWU of items in DD. The whole process in this phase can be divided into map stage and reduce stage.

- In map stage the Mapper outputs a key-value pair And it is fed in scattered data base with a transaction. In favor of all item in transaction.
- In reduce stage Mappers are fed to Reducers.

B. Up tree construction

It is given away that the tree-based structure for high utility item set mining applies the divide-and-conquer technique in mining process. Thus, the search space can be separated into smaller subspaces. By applying strategy DGN, the utilities of the nodes that are nearer to the origin of a global UP-Tree are further reduced. DGN is especially suitable for the databases containing lots of long transactions. In further words, the additional items a transaction contains, the more utilities can be discarded by DGN. On the contrary, conventional TWU mining representation is not suitable for such database since the transaction contains more items and then it has higher TWU. In subsequent subsections, we explain the procedure of constructing a global UP-Tree with strategies DGU and DGN.The global UP-Tree is completed with two database scans. In the primary scan, every transaction TU is computed and every 1- item's TWU is also accumulated. After first scan , it resulted wanted items and unwanted items. DGU is applied , transactions are reorganized by pruning the unwanted items and cataloging the remaining promising items in a fixed order. TWU order used for reorganized transaction. TWU downward order is mentioned the performance of this method. Then a function Insert Reorganized Transaction called to apply DGN during constructing a global UP-Tree.

C. Data transformation phase

In data transformation phase removes all low TWU 1-itemsets from DD and sorts remaining items in a TWU ascending order. Then, this algorithm transforms each reorganized transaction in DD into a special structure called u-transaction. The transaction utility of T' is denoted as TU and defined as the summary of utilities of every part of the items.

D. Mining phase

In mining phase, it discovers HUIs through several iterations. In the k-th iteration, all the HUIs of length k are discovered by performing a MapReduce pass.

E. Performance evaluation

The performance evaluation is conducted to prove the effectiveness of the proposed methodology in terms improved accuracy and reduced time complexity than the existing work. UP-Tree a compact tree structure, used make possible performance , keep away from scanning original database repeatedly and to preserve the information of transactions and high utility itemsets. Two strategies are applied to reduce the overvalued utilities stored in the nodes of global UP-Tree. In an UP-Tree, each node N contains name, count, node profit, parent node of N, node relation which points to a node whose item name is the same as name. A header table is employed to make traversal of UP-Tree. In header table maintains the values of item name, an overestimated value and a link which points to the final incidence of the node. So the nodes which contains same name can be traversed very efficiently. In the initial scan, all transaction's TU is computed and each 1- item's TWU is accumulated. After visiting all needed items, DGU is starts and transactions are updated by removing the unnecessary items. The required items sorted by TWU order.

Datasets

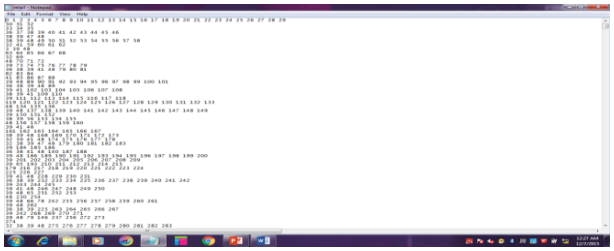

Fig. 2: Retail Datasets

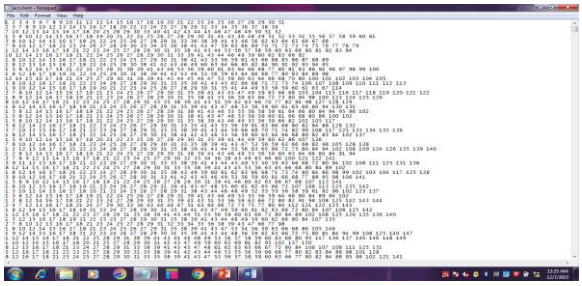

Figure 3: Accident Datasets

4. Conclusion

A PHUI pattern growth algorithm doesn't effectively prune the unpromising item sets for a large datasets. So a novel algorithm called UP-Growth is proposed . A UP-Tree is suggested for continuing the information of high utility item sets. By using the four strategies the search space and the number of scan are effectively reduced. This algorithm proposed to reduce the number of candidate itemsets and reduce the number of scans. PHUI can be efficiently generated from the UP-Tree with only two database scans. Moreover these algorithms can be used to decrease the overestimated utilities and enhance the performance of utility mining.

References

[1] Agrawal, R., and Srikant, R,"Fast algorithms for mining association rules",*20thVLDB Conference*,pp. 203-208, 1994.

[2] Gaber, M., Zaslavsky, A., Krishnaswamy, S,"Mining data streams: a review", *ACMSigmod Record* 34(2),pp. 18–26, 2005.

[3] Yao, H., Hamilton, H., Butz, C,"A foundational approach to mining itemset utilities from databases",In: *The 4th SIAM International Conference on Data Mining*, pp. 482–486, 2004.

[4] Liu .Y., Liao, W.K., Choudhary, A, "A two-phase algorithm for fast discovery of high utility itemsets", pp. 689–695, 2005.

[5] Hong, T.P., Lee, C.H., Wang, S.L,"Effective utility mining with the measure of average utility", *Expert Systems with Applications* 38,pp. 8259–8265, 2011.

[6] Lan, G.C., Hong, T.P., Tseng, V.S, "A Projection-Based Approach for Discovering High Average-Utility Itemsets", Journal of Information Science and Engineering 28, pp. 193–209.2012.

[7] Vo, B., Coenen, F., Le, B,"A new method for mining Frequent Weighted Itemsets basedon WIT-trees", *Expert Systems with Applications* 40, pp. 1256–1264, 2013.

[8] Fournier-Viger, P., Wu, C.-W., Zida, S., Tseng, V.S, "FHM :Faster High-Utility Itemset Mining using Estimated Utility Co-occurrence Pruning", In: *Springer, ISMIS 2014. LNCS*, vol. 8502,pp. 83–92, 2014.

[9] Tseng, V.S., Shie, B.-E., Wu, C.-W., Yu, P.S, "Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases", *IEEE Trans. Knowl. DataEng*. 25(8),pp. 1772–1786, 2013.

[10] Yin, J., Zheng, Z., Cao, L., Song, Y., Wei, W,"Efficiently Mining Top-K High Utility Sequential Patterns", *In Proceedings of ICDM 2013*, pp. 1259–1264, 2013.

[11] Baralis, E., Cerquitelli, T., &Chiusano, S,"IMine: Index support for item set mining", *IEEE TKDE Journal*, 21(4), 493–506, 2009.

[12] Han, J., Cheng, H., Xin, D., & Yan, X.," Frequent itemset mining: Current status and future directions", *DMKD Journa*l, 15(1), pp .55–86, 2007.

[13] Angel Latha Mary.S, Clement King.A "Comparing and Identifying Common Factors in Frequent Item set algorithms in Association Rule" in *IEEE proceedings on Computing, Communication and Networking,* Publication Date: 18-20 Dec. 2008, On page(s): 1-5, ISBN: 978-1-4244-3594-4,Digital Object Identifier: 10.1109/ ICCCNET. 2008. 4787769, Current Version Published: 2009-02-24.

[14]Angel Latha Mary.S, Shankar Kumar.K.R "Study Experiments on Frequent Itemset Algorithms in Association Rule" *International Journal of Data Mining Techniques* Volume: 01 No: 01 January 2012. pp34-41.